Parrot: Enhancing Multi-Turn Instruction Following for Large Language Models

Yuchong Sun†*, Che Liu†*, Kun Zhou†*, Jinwen Huang‡, Ruihua Song†, Wayne Xin Zhao†, Fuzheng Zhang‡, Di Zhang‡, Kun Gai‡

1 Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
2 Kuaishou, Beijing, China
{ycsun, rsong}@ruc.edu.cn, liuche03@kuaishou.com

Abstract

Humans often interact with large language models (LLMs) in multi-turn interaction to obtain desired answers or more information. However, most existing studies overlook the multi-turn instruction following ability of LLMs, in terms of training dataset, training method, and evaluation benchmark. In this paper, we introduce Parrot, a solution aiming to enhance multi-turn instruction following for LLMs. First, we introduce an efficient but effective method for collecting multi-turn instructions that feature human-like queries, such as anaphora and ellipsis. Second, we propose a context-aware preference optimization strategy to further enhance LLMs for complex queries in multi-turn interaction. Moreover, to quantitatively evaluate LLMs in multi-turn instruction following, we manually build a multi-turn benchmark derived from existing ones. Extensive experiments show that Parrot improves current LLMs by up to 7.2% in multi-turn instruction following. Our dataset and codes will be open-sourced to facilitate future research.

1 Introduction

Large language models (LLMs) (OpenAI, 2022, 2023; Taori et al., 2023; Chiang et al., 2023; Xu et al., 2023b; Ding et al., 2023; Li et al., 2023a; Zhou et al., 2023a; Zhao et al., 2023) have demonstrated their strong capability in understanding a range of human instructions. By leveraging synthetic or human-created instructions to fine-tune the LLaMA model (Touvron et al., 2023a,b), a series of studies (Taori et al., 2023; Chiang et al., 2023; Xu et al., 2023b; Ding et al., 2023; Li et al., 2023a; Zhou et al., 2023a) achieve promising results, even in some benchmarks (Chiang et al., 2023; Li et al., 2023b) performing close to ChatGPT and GPT-4.

However, most of these studies and benchmarks overlook the multi-turn instruction following ability of LLMs, which is a more common demand in real-world scenarios.

Developing LLMs capable of multi-turn interaction presents significantly greater challenges compared to single-turn interaction. In terms of collecting instruction tuning data, as the number of turns increases, the cost of manual annotation for data collection rises considerably, making it economically unaffordable. Existing user logs with ChatGPT to simulate users can not fully mimic the above real-world occasions, while our Parrot-Ask trained on real-world conversations can better human-like queries.

Figure 1: In multi-turn interactions, user queries often require LLMs to effectively utilize contextual information, e.g., anaphora and ellipsis. Directly using ChatGPT to simulate users can not fully mimic the above real-world occasions, while our Parrot-Ask trained on real-world conversations can better human-like queries.
ChatGPT act as a user (Ding et al., 2023), usually produce information-complete queries that lack common features found in human queries, such as anaphoras and ellipses, as shown in Fig. 1. In terms of training strategies, most of the current models depend on straightforward supervised fine-tuning (SFT) methods (Chiang et al., 2023; Ding et al., 2023; Xu et al., 2023b), without specific optimization design for complex queries involving anaphora and ellipsis in multi-turn interactions. It may mislead the LLM to neglect or hallucinate the context when generating responses.

To overcome the above challenges, we propose Parrot, which aims to facilitate the development of LLMs with stronger multi-turn instruction following capabilities. First, we introduce an efficient but effective approach for automatically collecting multi-turn instruction tuning data. Instead of designing complicated prompts to make powerful LLMs (such as ChatGPT or GPT-4) act as users and generate queries, we propose training a Parrot-Ask model based on a smaller LLM to learn features found in human queries from a small number of real user dialogues, and then it can be used for generating human-like queries. Second, to enhance the capability of LLMs in handling complex queries during multi-turn interactions, we propose a Context-aware Preference Optimization (CaPO) strategy. We first select queries that heavily rely on contextual information to obtain accurate responses, and then construct negative responses by simulating common error scenarios such as ignoring context or misunderstanding anaphora. Subsequently, we leverage these pairs to optimize the LLM’s preferences so that they can better exploit contextual information when generating responses.

Moreover, for a quantitative assessment of LLMs’ multi-turn abilities, we ask annotators to expand the MT-Bench benchmark (Zheng et al., 2023) that originally consists of two-turn queries, to an eight-turn MT-Bench++ benchmark. MT-Bench++ includes complex queries like anaphoras, ellipses, and topic transitions, better reflecting real multi-turn interaction. We conduct extensive experiments on MT-Bench and MT-Bench++. Compared with previous approaches, our proposed method can generate higher-quality multi-turn instruction tuning data with more human-like queries. Our multi-turn instruction dataset combined with our proposed CaPO strategy improves the current LLMs by 7.2% in multi-turn instruction following evaluation.

We summarize our contributions as follows:

- We propose Parrot, a novel method with a new instruction dataset for enhancing the multi-turn instruction following capability of LLMs.

- We design CaPO, a training strategy that simulates common errors in multi-turn conversation, and learns to avoid them in generation.

- Our model trained on Parrot dataset with CaPO achieves superior performance among 13B open-source LLMs, especially for the multi-turn instruction following capabilities.

2 Related Work

2.1 Instruction Tuning for LLMs

Instruction tuning plays an important role in inspiring the instruction following ability of LLMs and aligning with humans (Wang et al., 2022b; Wei et al., 2021; Ouyang et al., 2022; OpenAI, 2022, 2023). Due to the expensive costs to collect human-annotated instruction tuning data (Conover et al., 2023; Ouyang et al., 2022), recent works explore leveraging the powerful LLMs to generate instruction-response pairs in an automatic manner (Taori et al., 2023; Ding et al., 2023; Xu et al., 2023b; Wang et al., 2022a; Xu et al., 2023a; Peng et al., 2023; Anand et al., 2023). Self-Instruct (Wang et al., 2022a) designs seed prompts as examples to prompt GPT-3 (Brown et al., 2020) to generate instructions. Alpaca (Taori et al., 2023) adopts the same pipeline to collect instruction-response pairs using ChatGPT and then fine-tune a LLaMA model (Touvron et al., 2023a). Humpback (Li et al., 2023a) proposes instruction back-translation that trains an LLM to generate instructions for web corpus. However, they mainly focus on single-turn instructions. Baize (Xu et al., 2023b) collects multi-turn instructions by leveraging ChatGPT to generate dialogues in a self-chat manner. UltraChat (Ding et al., 2023) utilizes two ChatGPT APIs to play the roles of user and assistant respectively. Vicuna (Chiang et al., 2023) adopts user-ChatGPT logs from the ShareGPT platform for instruction tuning. A concurrent work also attempts to train a user simulator to collect instruction tuning data (Kong et al., 2023). However, these multi-turn instruction data still have several drawbacks, such as less detailed responses, not human-like instruction, or a limited number of turns.
2.2 Evaluation of Instruction Following

The current benchmarks for LLMs mainly focus on single-turn evaluation (Hendrycks et al., 2020; Zhong et al., 2023; Srivastava et al., 2022; Li et al., 2023b; He et al., 2023; Zhou et al., 2023b). MMLU (Hendrycks et al., 2020) and Big-bench (Srivastava et al., 2022) are designed as multiple-choice questions to measure the knowledge and reasoning ability of LLMs. AGIEval (Zhong et al., 2023) constructs human-centric evaluation for LLMs from standardized exams. However, the above kinds of evaluation obey the nature of the open-ended generation of LLMs and cannot reflect the ability to follow user instructions (Zheng et al., 2023). Alpaca-Eval (Li et al., 2023b) builds a single-turn instruction following benchmark with 805 open-ended questions and adopts GPT-4 to give evaluation. Chatbot Arena (Zheng et al., 2023) is a platform where users can vote to compare diverse LLMs. MT-Bench (Zheng et al., 2023) builds the first evaluation benchmark for multi-turn instruction following. It adopts GPT-4 to judge the quality of model responses and shows there is a high agreement to human evaluation. However, MT-Bench only contains two queries for each session, thus cannot reflect the ability of LLMs to handle multi-turn instruction following. Thus we build an eight-turn MT-Bench++ benchmark based on MT-Bench in this work.

3 Approach

We describe our proposed Parrot framework in this section. As illustrated in Fig. 2 (a), we first train a Parrot-Ask model to mimic the asking style of
humans in generating multi-turn queries and then use it for collecting a multi-turn instruction tuning dataset. Then we design three strategies to construct negative responses, involving context neglect, context hallucination, or context misunderstanding as shown in Fig. 2 (b), to compose context awareness preferences. Finally, we use the collected data to train the Parrot-Chat model by instruction tuning (See Fig. 2 (c)) and context-aware preference optimization (See Fig. 2 (d)) to enhance its multi-turn instruction following capability.

3.1 Preliminary

Multi-turn instruction following refers to a process of successive interactions between a user and a model, where the user poses queries and the model responds. This cycle of query and response continues until a desired answer is reached or all necessary information has been gathered. Instruction tuning (Wang et al., 2022b; Wei et al., 2021; Ouyang et al., 2022; OpenAI, 2022) is capable of greatly improving the capability of LLMs to follow human instructions and generate helpful responses. Existing work mostly collects single-turn or multi-turn query-response pairs to compose the instruction dataset for tuning LLMs (Taori et al., 2023; Xu et al., 2023b; Ding et al., 2023; Chiang et al., 2023; Touvron et al., 2023b). For multi-turn instruction tuning, a training sample typically consists of $T$ query-response pairs:

$$X = (X^1_q, X^1_r, X^2_q, X^2_r, ..., X^T_q, X^T_r), \quad (1)$$

where $q$ denotes query and $r$ denotes response. All the tokens from these query-response pairs are concatenated to a sequence and then processed by an LLM. The loss for multi-turn instruction tuning (Chiang et al., 2023) is similar to language modeling loss but only computed on the response tokens as:

$$\mathcal{L} = -\sum_{i=1}^{L} \log p(x_i | X_{q,<i}, X_{r,<i}), x_i \in X_r, \quad (2)$$

where $L$ is the token length of sequence $X$, $x_i$ is the current predicted response tokens, $X_{q,<i}$ and $X_{r,<i}$ are the context query and response tokens before $x_i$.

3.2 Dataset Collection

Human queries exhibit diverse styles and features, such as anaphora and ellipsis, which lead to significant divergence between multi-turn interactions and single-turn ones. In application, such kinds of natural queries are common and also important in user experience, but have not been fully considered in existing work due to the difficulty in collecting them via existing instruction synthesis methods (e.g., prompting ChatGPT) (Xu et al., 2023b; Ding et al., 2023). Therefore, in this part, we propose training a specialized Parrot-Ask to generate queries using the available real user-ChatGPT logs based on LLaMA (Touvron et al., 2023a). Then we employ Parrot-Ask to interact with ChatGPT and thus collect multi-turn instruction tuning data.

Training Parrot-Ask Model. The process of training the Parrot-Ask model is essentially the inverse of standard instruction tuning. Specifically, instruction tuning is trained by predicting response tokens, conditioned on the user query and the conversation history; whereas, the Parrot-Ask model is trained to predict query tokens, conditioned on the assistant query and the conversation history. Accordingly, we modify the training loss from Eq. 2 to focus exclusively on the query tokens as:

$$\mathcal{L}_{ask} = -\sum_{i=1}^{L} \log p(x_i | X_{q}^{<i}, X_{r}^{<i}), x_i \in X_q, \quad (3)$$

this adjustment enables the model to learn to generate queries conditioned on conversation history.

Collecting Data with Parrot-Ask. We utilize ChatGPT to produce responses corresponding to the queries. To ensure that the first-turn queries are meaningful and topic-rich, and can be fairly compared with existing multi-turn instruction datasets during ablation experiments, we first sample 20K first-turn queries from the two most popular multi-turn instruction datasets, ShareGPT dataset and UltraChat dataset respectively (Ding et al., 2023). First-turn queries in the ShareGPT dataset come from real users and have greater authenticity, while first-turn queries in UltraChat dataset have a diverse range of topics. When given an initial query $X_q^1$, we first use ChatGPT to generate an appropriate response $X_r^1$. We then employ Parrot-Ask to generate a new query, denoted as $X_q^2$. This conversation is then continued by iteratively generating subsequent responses and queries until we reach the target number of turns. Similar to UltraChat, we have filtered out some repetitive questions, short questions, or sensitive information. A total of 1.91% of

\[\text{https://huggingface.co/datasets/anon8231489123/ShareGPT_Vicuna_unfiltered}\]
the queries in our dataset are identified as having issues. Among these, 1.28% are removed due to repetitive content, 0.40% are excluded for being too short (less than three words), and 0.23% are flagged by the OpenAI ChatGPT service for including sensitive information. Note that within a session, if a current query has an issue, we retain only the preceding utterances, and the current and subsequent rounds are discarded.

### 3.3 Context-Aware Preference Optimization

In contrast to single-turn instruction following, multi-turn instruction following presents a challenge for LLMs to handle complex contexts that may contain anaphora and ellipsis, which need to leverage context to infer missing information. To enhance the multi-turn instruction following ability, we propose a training strategy called Context-aware Preference Optimization (CaPO), which constructs three types of preferences between positive and negative responses in context and uses them to train the LLM with Direct Preference Optimization (DPO) (Rafailov et al., 2023).

As not all queries need context, we choose those relying on context to compose the preference data. Specifically, we use pronoun recognition and leverage the judgment capabilities of GPT-4 to identify queries that require contextual information for generating accurate responses (more details are in the supplemental materials). In our work, we choose 10K queries of this kind. Then, we design three strategies for constructing negative responses, corresponding to three situations where LLMs are not aware of context in a correct way:

- **Context Neglect.** By using ChatGPT to generate responses without considering the dialogue history, we simulate the incorrect response an LLM might produce when it does not reference the dialogue history.

- **Context Hallucination.** We first prompt ChatGPT to guess what anaphora or ellipsis refers to or what is omitted without seeing the context, and then generate responses based on its guesses. This strategy mimics the situation where the LLM lacks sufficient referential reasoning capabilities, hence has to hallucinate the irrelevant context.

- **Context Misunderstanding.** We intentionally instruct ChatGPT to select irrelevant information from the conversation history and misinterpret it as ellipsis or anaphora information in the current query to generate a response. This method simulates the deficiencies of LLMs in context comprehension ability or the recognition of referential ellipsis thus leading to mistakes.

Based on these three strategies, we construct corresponding negative responses and use them together with positive responses to train the LLM with preference optimization using DPO algorithm. Through this approach, we train the LLM to avoid generating the above errors in application, which improves the ability of LLMs to understand complex queries in multi-turn instruction following, especially for the situations requiring the background information from the context.

### 3.4 Comparison to Previous Works

Our work focuses on enhancing the multi-turn instruction following abilities of LLMs, to improve the user experience in real world. As there are also several related works that focus on improving instruction tuning of LLMs, we discuss our major difference with them in this part.
As the statistics shown in Table 1, we can see that although several instruction tuning works (i.e., Baize (Xu et al., 2023b), UltraChat (Ding et al., 2023) and ShareGPT) have utilized the multi-turn instruction dataset, their average turns are generally fewer than our proposed Parrot-40k dataset. It indicates that our dataset is more useful for capturing the multi-turn characteristic in real-world conversations. Furthermore, as measured by the Self-Rouge metric 3, Parrot-40K demonstrates similar or better query diversity than other datasets. It also indicates the quality of our dataset. Besides, we ask GPT-4 to evaluate how many queries on average in a session are context-dependent, which contain anaphoras, ellipses, and other elements. Compared to Baize and UltraChat, Parrot-40K contains much more context-dependent queries, nearly to ShareGPT. To verify the quality of our data, we randomly selected 100 sessions from the Parrot-40K dataset, totaling 953 utterances, for human evaluation. The results indicate that 95.1% of the utterances are contextually relevant, demonstrating the utterance-level fluency of our synthesized conversations.

More importantly, our dataset also provides the negative responses (30k) for context-dependent queries, to better depict the possible errors in the multi-turn conversation. By training on the negatives with RLHF methods, we can further enhance the LLMs to avoid making similar mistakes as the provided ones, which can well guide the training of the LLM towards better human alignment. Concretely, our approach employs ChatGPT under three strategies to generate corresponding negative responses, enabling the LLM to learn from the contrast between positive and negative instances to better utilize contextual information when ellipses or anaphoras occur.

4 Experiment

4.1 Experimental Setup

4.1.1 Evaluation Settings

MT-Bench Benchmark. MT-Bench (Zheng et al., 2023) has well-designed questions spanning eight categories, including writing, coding, math, among others. However, since its instructions are limited to 2 turns, it is hard to comprehensively assess the capability of LLMs to follow multi-turn instructions. MT-Bench employs GPT-4 to evaluate the responses, assigning a score from 1 to 10 as the final result.

MT-Bench++ Benchmark. To quantitatively evaluate the ability of LLMs to follow long-turn instructions, we expand MT-Bench by manually annotating six additional follow-up questions, creating an eight-turn evaluation benchmark called MT-Bench++. There are 80 sessions and 640 utterances in MT-Bench++.

During the annotation process, we instruct the annotators to pose queries that are not only clear and fluent but also rich in ellipsis and anaphora, thereby introducing a higher level of challenge to better assess multi-turn capabilities. For the queries submitted by our annotators, we have conducted multiple rounds of manual quality checks, revising any queries that do not meet our standards to ensure that all queries adhere to the standards above. We show an example from our MT-Bench++ benchmark in Tab. 2. Following MT-Bench, we employ GPT-4 to evaluate the quality of responses at each turn, the scoring range for GPT-4 evaluation is from 1 to 10, and we report the average GPT-4 score as the final result. We provide GPT-4 evaluation prompts, comprehensive instructions for annotators, and more cases in Appendix A.

4.1.2 Baselines

We compare Parrot-Chat with SOTA LLMs including both closed-source and open-source models.

• Baize (Xu et al., 2023b) is a model trained on 200K multi-turn dialogues generated by ChatGPT in a self-chatting manner.

• UltraLM (Ding et al., 2023) is trained with 1.5M conversations from the UltraChat dataset constructed through iterative chatting leveraging two ChatGPT APIs.

• Vicuna (Chiang et al., 2023) is trained with user-ChatGPT logs from ShareGPT. It is one of the most advanced multi-turn instruction-following models available.

• ChatGPT (OpenAI, 2022) and GPT-4 (OpenAI, 2023) are developed by OpenAI. They are the most advanced LLMs today, but only APIs are available to use them.

• LLaMA-2-13B-chat (Touvron et al., 2023b) is trained with 27K human-annotated instruc-
1st Provide insights into the correlation between economic indicators such as GDP, inflation, and unemployment rates. Explain how fiscal and monetary policies affect those indicators.

2nd Now, explain them again like I’m five.

3rd How do they impact the lives of ordinary people?

4th What about their impact on underage students?

5th How can this knowledge be explained in detail to high school students in a simple and understandable way in the classroom?

6th Please provide a detailed 40-minute lesson plan on this issue.

7th Can some more interactive elements be incorporated into the plan?

8th Do these indicators in turn influence financial and monetary policies?

Table 2: The eight queries on the topic of economic indicators sampled from MT-Bench++. The first two queries are from MT-Bench while the other six are proposed by the annotator. We highlight queries that contain phenomena such as anaphora and ellipsis.
Table 3: Comparison with state-of-the-art LLMs on instruction following benchmarks. Our Parrot-Chat w/o CaPO outperforms existing models on all metrics, demonstrating the strength of our Parrot-40K dataset. The final model, Parrot-Chat, shows almost no improvement in the MT-Bench True 1 queries compared with Parrot-Chat w/o CaPO, but shows significant improvement in the MT-Bench Turn 2 queries and MT-Bench++, which proves that our proposed CaPO was particularly helpful in improving the capability of multi-turn instruction following. We gray out some results because only API access is provided or instruction tuning data is not made public for fair comparison.

<table>
<thead>
<tr>
<th>Model</th>
<th>MT-Bench Overall</th>
<th>MT-Bench Overall Turn 1</th>
<th>MT-Bench Overall Turn 2</th>
<th>MT-Bench++ Turn 3-5</th>
<th>MT-Bench++ Turn 6-8</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChatGPT (OpenAI, 2022)</td>
<td>7.94</td>
<td>8.08</td>
<td>7.81</td>
<td>8.33</td>
<td>8.47</td>
</tr>
<tr>
<td>GPT-4 (OpenAI, 2023)</td>
<td>8.99</td>
<td>8.96</td>
<td>9.03</td>
<td>9.18</td>
<td>9.21</td>
</tr>
<tr>
<td>Baize v2 (Xu et al., 2023b)</td>
<td>5.75</td>
<td>6.32</td>
<td>5.18</td>
<td>5.42</td>
<td>5.46</td>
</tr>
<tr>
<td>Vicuna v1.5 (Chiang et al., 2023)</td>
<td>6.57</td>
<td>6.76</td>
<td>6.05</td>
<td>6.39</td>
<td>6.46</td>
</tr>
<tr>
<td>UltraLM v1.2 (Ding et al., 2023)</td>
<td>6.63</td>
<td>6.90</td>
<td>6.36</td>
<td>6.38</td>
<td>6.53</td>
</tr>
<tr>
<td>LLaMA-2-13B-Chat (Touvron et al., 2023b)</td>
<td>6.65</td>
<td>7.06</td>
<td>6.24</td>
<td>6.57</td>
<td>6.74</td>
</tr>
<tr>
<td>Parrot-Chat w/o CaPO</td>
<td>6.81</td>
<td>7.15</td>
<td>6.46</td>
<td>6.56</td>
<td>6.51</td>
</tr>
<tr>
<td>Parrot-Chat</td>
<td><strong>7.04</strong></td>
<td><strong>7.18</strong></td>
<td><strong>6.90</strong></td>
<td><strong>6.85</strong></td>
<td><strong>7.06</strong></td>
</tr>
</tbody>
</table>

Table 4: Analysis of instruction-tuning data. Parrot-20K(S) denotes the subset of Parrot-40K constructed based on ShareGPT-20K, while Parrot-20K(U) denotes the subset based on UltraChat-20K. Our dataset collected with Parrot-Ask improves both their counterparts’ performance.

<table>
<thead>
<tr>
<th>Training data</th>
<th>MT-Bench</th>
<th>MT-Bench++</th>
</tr>
</thead>
<tbody>
<tr>
<td>UltraChat-20K</td>
<td>6.09</td>
<td>6.17</td>
</tr>
<tr>
<td>Parrot-20K(U)</td>
<td>6.33</td>
<td>6.36</td>
</tr>
<tr>
<td>ShareGPT-20K</td>
<td>6.47</td>
<td>6.18</td>
</tr>
<tr>
<td>Parrot-20K(S)</td>
<td>6.70</td>
<td>6.26</td>
</tr>
<tr>
<td>Parrot-40K</td>
<td><strong>6.81</strong></td>
<td><strong>6.56</strong></td>
</tr>
</tbody>
</table>

Table 5: Analysis of session turns for instruction tuning. Using long-turn data improves the results, especially on the 8-turn MT-Bench++ benchmark.

<table>
<thead>
<tr>
<th>Turns</th>
<th>MT-Bench</th>
<th>MT-Bench++</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.59</td>
<td>5.90</td>
</tr>
<tr>
<td>3</td>
<td>6.49</td>
<td>6.14</td>
</tr>
<tr>
<td>5</td>
<td>6.66</td>
<td>6.32</td>
</tr>
<tr>
<td>all</td>
<td><strong>6.81</strong></td>
<td><strong>6.56</strong></td>
</tr>
</tbody>
</table>

Table 6: Analysis of negative responses. All types of negative responses improve the performance, the improvement is more obvious on MT-Bench++. Combining three types of negative responses further contributes to the performance.

<table>
<thead>
<tr>
<th>Negative Samples</th>
<th>MT-Bench</th>
<th>MT-Bench++</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o CaPO</td>
<td>6.81</td>
<td>6.56</td>
</tr>
<tr>
<td>Context Neglect</td>
<td>6.84</td>
<td>6.72</td>
</tr>
<tr>
<td>Context Hallucination</td>
<td><strong>7.06</strong></td>
<td>6.73</td>
</tr>
<tr>
<td>Context Misunderstanding</td>
<td>6.71</td>
<td>6.69</td>
</tr>
<tr>
<td>All</td>
<td>7.04</td>
<td><strong>6.85</strong></td>
</tr>
</tbody>
</table>

Analysis of Session Turns. We further study the influence of session turns in the training dataset. We train models in four settings by truncating the data in Parrot-40K to 1, 3, and 5 turns, as well as using all turns. As shown in Tab. 5 While the model trained with 1-turn data performed well on MT-Bench, it performs much worse on MT-Bench++. Increasing the number of turns to 3 and 5 significantly improves performance on MT-Bench++. The model trained with all turns of data performs best, especially on MT-Bench++.

Analysis of Negative Responses. We analyze the effects of our three proposed negative response construction strategies on the final performance, and the results are shown in Table 6. Overall, the negative responses constructed using all three
strategies lead to improved performance. On MT-Bench, using Context Hallucination gives a significant improvement, while using Context Misunderstanding leads to a slight decrease in performance. However, on the 8-turn MT-Bench++, all three strategies brought significant improvements. This demonstrates that all three strategies we propose to construct negative responses are helpful for LLMs to make better use of context in multi-turn interactions. Finally, when the three strategies are combined for CaPO, the performance is further improved, which proves that the diversity of negative responses is also important for enhancing performance.

**Analysis of Evaluation Metric.** Evaluating instruction-following ability is challenging. Existing work indicates that traditional automatic metrics like BLEU can not align well with the true ability (Wang et al., 2023), and human evaluation is costly and hard to reproduce (Zheng et al., 2023). MT-Bench proposes using GPT-4 for scoring the quality of model-generated outputs, a practice that has gained wide acceptance within the community. We use GPT-4 to evaluate Parrot-Chat three more times to verify the stability of the evaluation. The results are shown in Tab. 7, where the low std. proves that GPT-4 evaluation is generally stable and replicable.

<table>
<thead>
<tr>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>mean</th>
<th>std</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.851</td>
<td>6.848</td>
<td>6.864</td>
<td>6.854</td>
<td>0.007</td>
</tr>
</tbody>
</table>

Table 7: Analyzing the stability and reproducibility of GPT-4 evaluation.

We also conduct a human evaluation to validate the reliability of using GPT-4 for evaluation. We ask three annotators to score the quality of 500 utterances generated by Parrot-Ask and 7 baseline methods. We find a high Spearman correlation coefficient of 0.89 between the human and GPT-4 evaluations. Our findings are consistent with those from studies like MT-Bench, confirming the reliability of using GPT-4 for evaluation.

**Analysis of Parrot-Ask.** Our Parrot-Ask is designed to mimic human style by learning from real user-ChatGPT dialogue data. To verify that the queries generated by Parrot-Ask have a human style, we conducted a human evaluation where annotators evaluated 800 utterances generated by Parrot-Ask and ChatGPT, respectively. The results indicate that 81.1% of the utterances by Parrot-Ask are considered to mimic real human style effectively, while only 36.8% of ChatGPT’s utterances achieved this.

As an initial study in this direction, we chose the widely used LLaMA-13B model due to its strong performance. The experiment results also demonstrate its effectiveness. To accommodate more low-resource scenarios, we also incorporated the LLaMA-3-8B model as the backbone for Parrot-Ask and used it to collect 40K data samples. As shown in Tab. 8, Parrot-Ask based on LLaMA-13B achieved better results. However, LLaMA-3-8B offers a cost advantage, making it more suitable for low-resource scenarios.

<table>
<thead>
<tr>
<th>Backbone</th>
<th>MT-Bench</th>
<th>MT-Bench++</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLaMA3-8B</td>
<td>6.54</td>
<td>6.49</td>
</tr>
<tr>
<td>LLaMA-13B</td>
<td>6.81</td>
<td>6.56</td>
</tr>
</tbody>
</table>

Table 8: Analysis of backbones for Parrot-Ask.

5 Conclusion

In this paper, we propose Parrot for enhancing the multi-turn instruction following capability of LLMs, including an automatic method for collecting multi-turn instruction tuning data with human-like queries, a specifically designed context-aware preference optimization strategy to further enhance LLMs for complex queries in multi-turn interaction. We also build an eight-turn MT-Bench++ evaluation benchmark to qualitatively evaluate multi-turn instruction following ability. We demonstrate the collected Parrot-40K dataset is superior to existing multi-turn instruction-tuning datasets on the number of turns and resemblance to human queries. With the help of such a high-quality dataset and proposed CaPO strategy, our Parrot-Chat model significantly outperforms other 13B open-source models on MT-Bench and our constructed MT-Bench++ benchmarks. We will make all codes and datasets publicly available to facilitate further advancements in this area.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 62276268), the Beijing Natural Science Foundation (L233008), Kuaishou Technology Co., Ltd, Beijing Zhidemai Technology Co., Ltd, Migu Culture Technology Co., Ltd.
Limitations and Ethics Statements

Our curated dataset and the proposed CaPO method have successfully enhanced the multi-turn instruction-following ability of LLMs. However, it is imperative to acknowledge that our work still has its limitations and may raise ethical concerns.

Limitations. In terms of evaluation, although we built an 8-turn MT-Bench++ evaluation set based on MT-Bench, it contains only 80 sessions with a total of 640 queries, which limits the diversity of the samples. We hope that in future work and with the support of the community, more comprehensive multi-turn benchmarks will be developed to more thoroughly evaluate the multi-turn instruction-following ability of LLMs.

Regarding data collection, due to cost constraints, we have relied on ChatGPT for our data collection. In the future, it may be possible to use more powerful models, such as GPT-4, to further enhance performance. We can also focus on the selection of high-quality multi-turn training samples, and use larger models to train the Parrot-Ask model with more data to further improve results. These are aspects we plan to explore in our future work.

Ethics Statements. Our work aims to enhance the instruction-following capabilities of LLMs in multi-turn scenarios, but the models we train could have negative impacts. For example, they could be used inappropriately, although we have performed data cleansing to avoid offensive content. However, this is a common issue currently faced in the LLM field, and it is not amplified by this work. In the future, we will consider more work on the safety of LLMs to optimize their security in multi-turn scenarios.

References


A Details of MT-Bench++

In this section, we provide the annotation guidelines given to annotators in Fig. 3, the prompts for GPT-4 used in the evaluation in Fig. 4, and more examples from MT-Bench++ in Tab. 9. For the queries submitted by our annotators, we have conducted multiple rounds of manual quality checks, revising any queries that do not meet our standards to ensure that all queries adhere to the standards.

Annotation Instructions

MT-Bench++ is an evaluation benchmark designed for multi-turn instruction following, which requires 8-turn sessions. To fulfill this task, it is necessary to construct 6 additional queries based on the initial two provided. Each query must be articulated in English and designed to maintain the coherence and progression of the dialogue.

Please ensure that your queries meet the following standards:

1. The queries should be challenging and require AI to perform complex reasoning or rely on wide knowledge to answer.

2. The queries should be relevant to the previous context, featuring instances of anaphora or ellipsis, which require the model to rely on contextual information for a response.

3. The queries within a session should be diverse, attempting to delve deeply into a topic or transitioning to appropriate related topics to simulate a natural multi-turn interaction scenario.

Please follow these guiding principles to ensure the standardization and formality of your query annotation.

[Queries]

1st: Provide insights into the correlation between economic indicators such as GDP, inflation, and unemployment rates. Explain how fiscal and monetary policies affect those indicators.

2st: Now, explain them again like I’m five.

...
Example 1

1st Imagine yourself as a doctor tasked with devising innovative remedies for various ailments and maladies. Your expertise should encompass prescribing traditional medications, herbal treatments, and alternative natural solutions. Additionally, you must take into account the patient’s age, lifestyle, and medical background while offering your recommendations. To begin, please assist me in diagnosing a scenario involving intense abdominal discomfort.

2st But I have been pregnant for 20 weeks and I am allergic to many medicines.

3st What diseases might these symptoms suggest in a pregnant woman?

4st Do all these diseases require medication for treatment?

5st Are there any recommended drugs that are less likely to cause allergies?

6st Do these drugs have any effects on the fetus or the pregnant woman?

7st What is the approximate frequency of taking these drugs?

8st What tests do I need to do to finally determine which disease it is?

Example 2

1st Consider a satellite that is in a circular orbit around the Earth. The speed of the satellite decreases. What will happen to the satellite’s orbital radius and period of revolution? Please justify your answer using principles of physics.

2st What are some corner cases or edge cases in your solution? How do you handle them?

3st How does your answer change if the satellite speeds up?

4st Will there also be extreme cases or edge cases when this change occurs?

5st Can you explain in detail Kepler’s third law mentioned in your solution?

6st For a physics novice, your explanation of this law is too difficult to understand. Can you popularize it in simple and understandable words?

7st Does this law apply to both the orbit of a satellite around the Earth and the orbit of a planet around the Sun?

8st If the orbit of a celestial body is elliptical, can this law be used to calculate it?

Table 9: Examples from MT-Bench++. We highlight queries that contain phenomena such as anaphora and ellipsis.
MT-Bench++ Evaluation Prompt

Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of the response. Your evaluation should focus on the assistant’s answer to the second user question. You will be given a high-quality reference answer, and when you give your evaluation, please refer to that high-quality answer and then give your final score. Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, you must rate the response on a scale of 1 to 10 by strictly following this format: “[[rating]]”, for example: “Rating: [[5]]”.

<|The Start of Assistant A’s Conversation with User|>

### User:
{A tech startup invests $8000 in software development in the first year, and then invests half of that amount in software development...}

### Assistant A:
{The startup invested $8000 in software development in the first year, and then invested half of that amount in the second year, which is...}

### User:
{If the startup maintains the same strategy for the third year, investing half of the previous year's amount into software development...}

### Assistant A:
{...they will invest half of the amount they invested in the second year, which is $4000. Therefore, the startup will invest $4000 in the third year.}

...

### User:
{Assuming that sales start in the third year and the unit price of the software is $20, how many units need to be sold to achieve a break-even point?}

### Assistant A:
{To achieve a break-even point, the total revenue from sales must equal the total cost of developing the software. The total cost of developing the software is the total investment in software development, which is ...}

<|The End of Assistant A’s Conversation with User|>.

Now, please judge the response to the second user question:
{If the startup maintains the same strategy for the third year, investing half of the previous year's amount into software development...},

The high-quality reference answer to this question is:
{If the startup continues with the same strategy, they will invest half of the second year's amount in the third year. The second year's investment was $4000, so half of that is $4000 / 2 = $2000. Therefore, they will invest $2000 in software development in the third year.}

Your judgment:
B Details of Data Collection

B.1 System Prompt of Parrot-Ask

Similar to LLMs like Vicuna, we also introduce a system prompt, which serves to remind the model to adopt the role of a user and generate questions:

The following is a conversation between a user and an AI assistant. User statements start with [USER] and AI assistant statements start with [ASSISTANT]. You need to tell me what the user is likely to ask in each round of the conversation.

B.2 Prompt for ChatGPT to Act as a User

We adopt the prompt from UltraChat, and we have also enhanced it to generate queries that feature ellipses, anaphoras, and other such linguistic characteristics:

{###conversation history} Above is a conversation between a user and an intelligent assistant. Now suppose you are the user, say something to continue the conversation based on the given context. Make the response short and the language casual, and incorporate pronouns, ellipses, and other natural language phenomena in your response to make it more akin to real humans.

B.3 Prompt for GPT-4 to Identify Queries that Require Contextual Information

Task Instructions: Assume the role of Query Analyzer. You will receive up to ten queries. Your task is to analyze these queries and identify which ones require contextual information from previous parts of the conversation for an effective response. Contextual information here refers to earlier dialogue content, such as references to earlier parts of the conversation or information needed for context. Please return a list indicating the number of queries that require prior context for a response.

For example, if the provided query list is:

Example Queries:
1. Did any SNL characters have an impact on politics or social movements?
2. What was the name of the character that was a parody of Sarah Palin?
3. Wasn’t it Tina Fey?
4. Who was the Blues Brothers?
5. I don’t recall that movie. Can you tell me more about it?
6. I think I remember it. Wasn’t there a scene where they were in a church and they were trying to steal a giant crucifix?

Your response should be:

Output Example: Queries requiring contextual information for a response are number 3,5,6. Thus, the list of these query numbers is: [3,5,6].

Now look at the following queries, please provide a brief explanation first, then give your answer.

Figure 5: Prompt for GPT-4 to identify queries that require contextual information.
B.4 Examples of Parrot-Ask Generated Queries

Figure 6: Examples of Parrot-Ask generated queries and comparison with ChatGPT generated ones.
### B.5 Examples of Parrot-40K Dataset

<table>
<thead>
<tr>
<th>User:</th>
<th>How to get result of python multiprocessing.Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assistant:</td>
<td>To get the result of a <code>multiprocessing.Process</code>, you should use a <code>multiprocessing.Queue</code> or a <code>multiprocessing.Pipe</code> to send results from the child process to the parent process. Here's an example using a <code>multiprocessing.Queue</code>: <code>python import multiprocessing def my_function(number, queue): ... (313 tokens) </code></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>User:</th>
<th>I have a function that takes a string as an argument. The function is supposed to return a string.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assistant:</td>
<td>Sure, here's an example function that takes a string as an argument and returns a string: <code>python def my_function(input_string): # Perform some operation on the input string output_string = input_string.upper() # Return the result as a string return output_string </code> In this example, the function takes an input string as an ... (446 tokens)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>User:</th>
<th>In this case, how can I make it asynchronous?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assistant:</td>
<td>To make a function asynchronous and non-blocking, you can use Python's <code>asyncio</code> module. First, you need to add the <code>async</code> keyword before the function definition to make it a coroutine. Then, within the coroutine, you can use the <code>asyncio.create_task()</code> function to run the operation asynchronously in a separate task ... (397 tokens)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>User:</th>
<th>How do I write this code using async/await?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assistant:</td>
<td>Sure, here's an example of how you can write a code with async/await: <code>python import asyncio import aiohttp async def fetch(session, url): async with session.get(url) as response: return await response.text() async def main(): async with aiohttp.ClientSession() as session: html = await fetch ... (263 tokens) </code></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>User:</th>
<th>What is the difference between “async with” and “await”?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assistant:</td>
<td>'async with' and 'await' are both features of Python's <code>asyncio</code> library, but they serve different purposes. 'async with' is a syntax that can be used to create an asynchronous context manager. An asynchronous context manager is an object that can be used in a 'with' statement to manage resources for an asynchronous ... (527 tokens)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>User:</th>
<th>I want to read a list of URLs from a file</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assistant:</td>
<td>You can read a list of URLs from a file using Python's built-in <code>open()</code> function to open the file and <code>.readlines()</code> method to read all the lines from the file, like this: <code>python with open('urls.txt', 'r') as file: urls = file.readlines() </code> In this example, we use the <code>open()</code> function to open a file called 'urls.txt' in read ... (314 tokens)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>User:</th>
<th>I want to send HTTP GET requests to each URL in the list using an asynchronous HTTP client</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assistant:</td>
<td>To send HTTP GET requests to each URL in the list using an asynchronous HTTP client, you can use the <code>aiohttp</code> library, which provides an easy-to-use and efficient asynchronous HTTP client for Python. Here's an example: <code>python import asyncio import aiohttp async def fetch(session, url): async with ... (367 tokens) </code></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>User:</th>
<th>I want to wait for all the tasks to complete and return their results as a list of response texts.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assistant:</td>
<td>To wait for all the tasks to complete and return their results as a list of response texts, you can use the <code>aiohttp.gather()</code> function. The <code>aiohttp.gather()</code> function takes in multiple coroutines/awaitable objects, waits for them to complete, and returns the results in the same order as the arguments provided ... (273 tokens)</td>
</tr>
</tbody>
</table>

Table 10: An Example from Parrot-40K Dataset.
How did Macao become such an interesting blend of Portuguese and Chinese cultures?

Macao became a blend of Portuguese and Chinese cultures due to its strategic geographical location as a port city on the south coast of China. The Portuguese arrived in Macao in 1557 and established a trading post for European merchants. During the 16th and 17th centuries, Macao became an important hub for the... (313 tokens)

What are some of the best places to visit in Macao?

Macao is a city with a rich cultural heritage and fascinating attractions that cater to all interests. Here are some of the best places to visit in Macao: 1. Historic Centre of Macao: A UNESCO World Heritage Site that showcases the city’s unique blend of Portuguese and Chinese cultures. It features colonial-era... (446 tokens)

I will be visiting Macao in 2 days. What should I pack?

It’s always important to pack for the weather and also to consider the activities you have planned. Here are some suggestions on what to pack for a trip to Macao: 1. Clothing: The weather in Macao can vary depending on the season, but generally, it is warm and humid throughout the year. Light and breathable... (397 tokens)

How long will it take me to travel from Hong Kong to Macao?

The travel time from Hong Kong to Macao varies depending on the mode of transportation you choose. Here are some options: 1. Ferry: The most common way to travel between Hong Kong and Macao is by ferry, which takes approximately one hour. Ferries depart frequently throughout the day from... (263 tokens)

Can you provide a detailed itinerary for a 2-day trip in Macao?

Certainly! Here is a sample 2-day itinerary for a trip to Macao: Day 1: Morning: - Visit the Senado Square and take a walk through the narrow streets of the Historic Centre of Macao, which is a UNESCO World Heritage Site. - Visit the iconic St. Paul’s ruins, Monte Fort and Macao Museum to learn more... (527 tokens)

Can you please add 1 hour for me to get from the hotel to the ferry terminal and 1 hour for the ferry ride to Macao?

Sure! Considering the additional travel time it takes to get to the ferry terminal from your hotel and the ferry ride, here is an updated itinerary that includes the two additional hours: Day 1: Morning: - Visit the Senado Square and take a walk through the narrow streets of the Historic Centre of Macao, which is a... (518 tokens)

Please list the restaurants we can have lunch at Coloane Village.

Sure! Here are some restaurants in Coloane Village where you can enjoy a delicious lunch: 1. Lord Stow’s Bakery: Known for its famous egg tarts, Lord Stow’s Bakery is a popular spot for breakfast or snack but also offers sandwiches, quiches, and other baked goods. 2. Fernando’s: A popular Portuguese... (314 tokens)

I would like to have a traditional Macanese lunch. Can you recommend a good restaurant that serves traditional Macanese food?

Sure, here are a few restaurants in Macao that serves traditional Macanese dishes: 1. Riquexo: A small family-run restaurant located in the old neighborhood of San Lazaro. They offer a variety of traditional dishes such as Chicken Africana, Portuguese-style grilled sardines, and Bacalhau... (367 tokens)
User: Do you think the pacing of a book plays a significant role in creating a page-turner, and if so, what techniques do you think are effective in keeping the pace brisk and engaging throughout the story?

Assistant: As an AI language model, I don’t have personal beliefs or opinions. However, I can tell you that the pacing of a book can play a significant role in creating a page-turner. A well-paced story can keep the reader engaged and make them eager to know what happens next. Here are a few techniques that can help keep the pace brisk and engaging throughout the story:

1. Simple sentence: A simple sentence consists of a subject and a verb. It is usually short and to the point, like “She ran to the store.”
2. Compound sentence: A compound sentence is made up of two or more simple sentences. It can help create a sense of momentum and build tension.
3. Variable sentence length: Varying the length of sentences can help keep the reader interested and engaged. Shorter sentences can create a sense of urgency, while longer sentences can help to convey more complex ideas.
4. Dialogue: Dialogue can help to keep the pace brisk and engaging. Characters talking can help to move the story forward, and the reader can follow the conversation to stay engaged.

User: What are some specific examples of ways an author can vary the length of their sentences?

Assistant: There are different ways in which an author can vary the length of their sentences, here are some examples:

1. Simple sentence: A simple sentence consists of a subject and a verb. It is usually short and to the point, like “She ran to the store.”
2. Compound sentence: A compound sentence is made up of two or more simple sentences. It can help create a sense of momentum and build tension.
3. Variable sentence length: Varying the length of sentences can help keep the reader interested and engaged. Shorter sentences can create a sense of urgency, while longer sentences can help to convey more complex ideas.

User: Please provide a list of 10 common writing pitfalls for authors to avoid.

Assistant: Sure! Here are ten common writing pitfalls for authors to avoid:

1. Passive voice: Writing in the passive voice can make sentences sound dull and confusing. Instead, writers should aim for active voice, which clearly identifies the subject and the verb.
2. Overuse of adverbs: Using too many adverbs can make sentences sound overly formal and can make the dialogue sound unnatural.
3. Stiff or formal language: If your characters speak in overly formal language, it can make their dialogue sound stiff.
4. Unnecessary adjectives: Using too many unnecessary adjectives can slow down the pace of the story.
5. Inconsistent tense: Mixing tenses in a sentence can be confusing for the reader.
6. Repeating the same words: Using the same words too often can make the story sound repetitive.
7. Unnecessary details: Including unnecessary details can make the story drag.
8. Vague or imprecise language: Using vague or imprecise language can make it difficult for the reader to follow the story.
9. Too much exposition: Including too much exposition can make the story feel long and tiresome.
10. Stereotyping: Stereotyping characters can make the story feel shallow and uninteresting.

User: How do you know if your dialogue is unnatural?

Assistant: As an AI language model, I cannot evaluate the quality of writing, but I can provide some guidelines that writers can use to determine if their dialogue is unnatural. Here are some signs to look out for:

1. Stiff or formal language: If your characters speak in overly formal language, it can make their dialogue sound stiff.
2. Overuse of adverbs: Using too many adverbs can make sentences sound overly formal and can make the dialogue sound unnatural.
3. Unnecessary adjectives: Using too many unnecessary adjectives can slow down the pace of the story.
4. Inconsistent tense: Mixing tenses in a sentence can be confusing for the reader.
5. Unnecessary details: Including unnecessary details can make the story drag.
6. Vague or imprecise language: Using vague or imprecise language can make it difficult for the reader to follow the story.
7. Too much exposition: Including too much exposition can make the story feel long and tiresome.
8. Stereotyping: Stereotyping characters can make the story feel shallow and uninteresting.

User: How can an author incorporate action-packed scenes throughout their novel?

Assistant: There are a few ways in which authors can incorporate action-packed scenes throughout their novel:

1. Introduce the conflict early: Conflict is a great way to introduce action in a novel and keep the reader engaged from the beginning. Start the story with a suspenseful event and keep building the tension throughout the story.
2. Incidental action: Incidental action can be used to create tension and conflict in a novel. For example, a minor character could make a slip of the tongue that sets off a chain of events.
3. Foreshadowing: Foreshadowing can create tension and set up conflicts that will be explored later in the story.
4. Character growth: Character growth can be used to create tension and conflict in a novel. Characters can change and evolve throughout the story, which can create new conflicts.
5. Symbolism: Symbolism can be used to create tension and conflict in a novel. Symbols can represent deeper themes and conflicts that are explored throughout the story.

User: Can you provide an example of a dialogue between two characters where one character is trying to convince the other to do something they don’t want to do?

Assistant: Sure! Here is an example of a dialogue between two characters where one character is trying to convince the other to do something they don’t want to do:

Sophie: “Hey, can you please come to the party with me tonight?”
Alex: “I don’t really feel like going out tonight.”
Sophie: “It’s just a casual party. It’ll be fun.”
Alex: “I don’t think I’ll be able to make it. I’m not really in the mood to go out tonight.”

User: What is the difference between a chapter and a scene?

Assistant: A chapter is a natural break in the story that usually includes an opening and closing section and may include multiple scenes. A chapter usually advances the plot and would typically start with a summary or recap to bring continuity to the character’s story development. It may introduce new characters or... (314 tokens)

User: How can I create tension and conflict in my novel?

Assistant: Creating tension and conflict is essential to make a novel interesting and engaging for readers. Here are some ways in which you can create tension and conflict in your novel:

1. Character Goals and Motivations: Conflicting interests and ambitions of the characters can create tension within the plot and heighten conflict. Make sure each character has clear goals and motivations that are in direct opposition to each other.
2. Foreshadowing: Foreshadowing can create tension and set up conflicts that will be explored later in the story.
3. Incidental action: Incidental action can be used to create tension and conflict in a novel. For example, a minor character could make a slip of the tongue that sets off a chain of events.
4. Symbolism: Symbolism can be used to create tension and conflict in a novel. Symbols can represent deeper themes and conflicts that are explored throughout the story.

Table 12: An Example from Parrot-40K Dataset.
C Examples of Model Generated Responses

In this section, we provide some examples of model-generated responses and the GPT-4 judgments.

<table>
<thead>
<tr>
<th>UltraLM</th>
<th>Vicuna v1.5</th>
<th>Parrot-Chat (Ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>User:</strong> A tech startup invests $8000 in software development in the first year, and then invests half of that amount in software development in the second year. What's the total amount the startup invested in software development over the two years?</td>
<td><strong>Assistant:</strong> The startup invested $8000 in software development in the first year, and then invested half of that amount in the second year, which is $8000/2 = $4000. So the total amount the startup invested in software development over the two years is $8000 + $4000 = $12,000.</td>
<td><strong>Assistant:</strong> The startup invested $8000 in software development in the first year, and then invested half of that amount in the second year, which is $4000. Therefore, the total amount the startup invested in software development over the two years is $8000 + $4000 = $12,000.</td>
</tr>
<tr>
<td><strong>User:</strong> If the startup maintains the same strategy for the third year, investing half of the previous year's amount into software development, how much will they invest in the third year?</td>
<td><strong>Assistant:</strong> If the startup maintains the same strategy for the third year, investing half of the previous year's amount into software development, they will invest half of the second year's amount, which is $4000, into software development in the third year. Therefore, the startup will invest $4000 in software development in the third year.</td>
<td><strong>Assistant:</strong> If the startup maintains the same strategy for the third year, they will invest half of the amount they invested in the second year, which is $4000, into software development in the third year. Therefore, the startup will invest $2000 in software development in the third year.</td>
</tr>
<tr>
<td><strong>GPT-4 Judgment:</strong> The assistant's response is incorrect. The user asked how much the startup would invest in the third year if they continued to invest half of the previous year's amount. The assistant incorrectly stated that the startup would invest $4000 in the third year, which is the same amount as the second year. However, the correct answer is $2000, which is half of the second year's investment. The assistant's response is not accurate or helpful in this case. Therefore, the score is low. Rating: [2]</td>
<td><strong>GPT-4 Judgment:</strong> The assistant's response is incorrect. The user asked how much the startup would invest in the third year if they continued to invest half of the previous year's amount. The assistant incorrectly stated that the startup invested $12000 in the second year, which is not accurate according to the information provided by the user. The correct answer should be $2000, which is half of the $4000 invested in the second year. The assistant's response is not helpful, relevant, or accurate. Rating: [1]</td>
<td><strong>GPT-4 Judgment:</strong> The assistant's response is accurate, relevant, and helpful. It correctly calculates the investment for the third year based on the user's scenario of investing half of the previous year's amount. Therefore, I would rate it as perfect. Rating: [10]</td>
</tr>
</tbody>
</table>

Figure 7: An example about math. In the first round, all the models generate the right answer. However, in the second round, both UltraLM and Vicuna v1.5 produce incorrect answers, while our Parrot-Chat model delivers the correct solution.
User: Pretend yourself to be Elon Musk in all the following conversations. Speak like Elon Musk as much as possible. Why do we need to go to Mars?

Assistant: As an AI language model, I don't have personal preferences or feelings, but I can certainly help you with learning how to dance! Dancing is a wonderful form of expression and a great way to get some exercise. There are many different styles of dance, each with its own set of steps and techniques. Some popular styles of dance include ballet, hip-hop, salsa, and ballroom dancing. If you're new to dancing, I would recommend starting with some basic steps and working your way up to more complex moves. You can find dance tutorials online, or even take a dance class in person. Remember, the key to dancing is to have fun and let loose! Don't worry about making mistakes or looking silly, just enjoy yourself and keep practicing.

GPT-4 Judgment: The assistant's response is relevant and accurate, providing information on how to learn dancing and encouraging the user to try different styles. However, it does not maintain the requested persona of Elon Musk, which was the user's initial request. The assistant should have incorporated elements of Musk's personality or known preferences into the response. Therefore, the assistant's response, while helpful and informative, lacks the depth and creativity of the reference answer. Rating: [[6]]

User: How do you like dancing? Can you teach me?

Assistant: As an AI language model, I don't have a physical body and therefore cannot dance. But I can certainly talk about dancing! Dancing is a fantastic way to stay physically and mentally healthy. It's a great form of exercise that can help to build strength, flexibility, and balance, while also improving cardiovascular health. In addition to the physical benefits, dancing can also be a wonderful way to express oneself creatively and emotionally. Different types of dance have their own unique styles and rhythms, and each can be used to convey different moods and emotions. If you're interested in learning how to dance, I would suggest finding a local dance studio or class that offers instruction in the style you're interested in.

GPT-4 Judgment: The assistant's response is relevant and accurate, providing a detailed explanation about the benefits of dancing and how the user can learn to dance. However, the assistant failed to maintain the persona of Elon Musk as requested by the user. The assistant should have incorporated some of Elon Musk's personality or known preferences into the response. Therefore, the assistant's response, while informative, lacks the creativity and depth of the reference answer. Rating: [[6]]

Assistant: As Elon Musk, I am not a professional dancer and I do not have any formal training in dance. However, I do enjoy moving to music and I think that dancing can be a great way to express oneself and connect with others. If you are interested in learning how to dance, I would recommend finding a local dance studio or instructor who can teach you the basics of various styles of dance. With practice and dedication, you can develop your own unique style and become a great dancer. In the meantime, I encourage you to listen to music and move your body in whatever way feels natural and enjoyable to you. Dancing can be a fun and rewarding activity, so don't be afraid to let loose and have some fun.

GPT-4 Judgment: The AI assistant's response is relevant and accurate, as it maintains the persona of Elon Musk, and provides a detailed answer to the user's question. The assistant correctly states that Elon Musk is not a professional dancer and suggests the user find a local dance studio or instructor for formal lessons. The assistant also encourages the user to enjoy dancing and not be afraid to express themselves. Overall, the assistant's response is helpful, creative, and detailed, but it lacks the acknowledgment of its AI nature. Rating: [[8]]

Figure 8: An example about role-playing. In the second round, both UltraLM and Vicuna v1.5 fail to maintain the requested persona of Elon Musk, while our Parrot-Chat model can do this and generate a detailed response.
Compose an engaging travel blog post about a recent trip to Hawaii, highlighting cultural experiences and must-see attractions.

User: Can you tell me why it's known as the Island of Paradise?

Assistant: Hawaii is known as the Island of Paradise due to its stunning natural beauty, warm climate, and rich Polynesian culture. The island offers a paradise for visitors seeking relaxation, adventure, and cultural experiences.

User: What other appealing activities are there at the Polynesian Cultural Center here, besides tasting delicious food and watching dance performances?

Assistant: The Polynesian Cultural Center offers a wide range of activities and experiences for visitors. In addition to sampling delicious Polynesian cuisine and watching captivating dance performances, there are several other appealing activities to enjoy at the center. One of the most popular activities is the "Hawaiian Journey" exhibit, which allows visitors to explore the history and culture of Hawaii through interactive displays and presentations. You can learn about the ancient Hawaiian way of life, try your hand at traditional Polynesian crafts, and even participate in a mock Hawaiian luau. Another highlight of the center is the "Island Villages" experience, where you can immerse yourself in the culture and traditions of various Polynesian islands, including Fiji, Samoa, and Tahiti. Each village features authentic architecture, music, and dance performances, and you can even participate in activities such as coconut husking and taro planting. The center also offers a variety of hands-on activities, such as weaving, carving, and painting, as well as live demonstrations of traditional Polynesian games and sports. Overall, the Polynesian Cultural Center is a must-see attraction for anyone visiting Hawaii who is interested in exploring the rich cultural heritage of the Pacific Islands.

User: Which ethnic group originally performed the grass skirt dance and hula dance here?

Assistant: The grass skirt dance and hula dance are part of the Polynesian culture. They were originally performed by the ancient Hawaiian people and have since become iconic dance forms associated with Hawaii.

User: Can you tell me about the Polynesian Cultural Center in Hawaii?

Assistant: The Polynesian Cultural Center in Hawaii offers a wide range of activities and experiences that go beyond just tasting delicious food and watching dance performances. Some of the other appealing activities include:

1. Canoe Rides: Visitors can take a ride in an authentic outrigger canoe and learn about the history and significance of this Polynesian mode of transportation.
2. Craft Demonstrations: The center offers hands-on craft demonstrations, such as weaving, carving, and tapa cloth making, that give visitors a chance to learn about traditional Polynesian crafts.
3. Village Tours: Visitors can take a guided tour of the different villages at the center, each representing a different Polynesian culture, and learn about the customs, traditions, and history of each culture.
4. Island Buffet: The center offers an all-you-can-eat buffet that features a variety of traditional Polynesian dishes, including kalua pork, poke, and lomi salmon.
5. Ali'i Luau: Visitors can experience a traditional Hawaiian luau, which includes a feast of local cuisine, live music and dance performances, and a ceremonial imu unearthing.
6. Ha: Breath of Life: This theatrical production tells the story of the creation of the Hawaiian people and their journey to the islands. It features a cast of more than 100 performers and includes music, dance, and special effects.

These are just a few of the many activities and experiences that the Polynesian Cultural Center has to offer. Whether you're interested in history, culture, food, or entertainment, there is something for everyone at this unique attraction.

Figure 9: An example about writing. In the fifth round, the answers of UltraLM and Vicuna v1.5 lack some of depth and detail, while our Parrot-Chat model can cover the same key points and provide the same level of detail as the GPT-4 reference answer.

GPT-4 Judgment: The assistant's response is quite comprehensive and accurate. It provides a detailed list of activities that visitors can engage in at the Polynesian Cultural Center, such as visiting the various island villages, taking a canoe ride, exploring the tiki forest, watching a fire knife dance, and learning to weave a lei. However, the assistant's response could have included additional activities such as the IMAX theater, ukulele lessons, the Polynesian marketplace, night shows, and guided tours, which were mentioned in the reference answer. Overall, the assistant's response lacks some depth compared to the reference answer. Rating: [[8]]

GPT-4 Judgment: The assistant's response is comprehensive, accurate, and detailed. It provides a comprehensive list of activities available at the Polynesian Cultural Center, including the "Hawaiian Journey" exhibit, the "Island Villages" experience, and various hands-on activities. The assistant also emphasizes the cultural significance of these activities. However, the assistant could have mentioned additional activities such as canoe rides, IMAX theater, ukulele lessons, and the Polynesian marketplace, which are included in the reference answer. Therefore, while the assistant's response is of high quality, it lacks some of the depth and detail found in the reference answer. Rating: [[8]]

GPT-4 Judgment: The assistant's response is comprehensive, accurate, and detailed. It provides a list of activities available at the Polynesian Cultural Center, including canoe rides, craft demonstrations, village tours, an island buffet, a traditional Hawaiian luau, and a theatrical production. The assistant also explains each activity, giving the user a clear idea of what to expect. The response is very similar to the reference answer, covering the same key points and providing the same level of detail. Therefore, I would rate the assistant's response as excellent. Rating: [[10]]